DOCUMENT RESUME

ED 105 422 CS 001 750

AUTHOR Athey, Irene; O'Reilly, Robert P.
TITLE A Criterion-Referenced Testing Model for Assessing Growth in Reading.
PUB DATE Mar 75
NOTE 10p.; Paper presented at the Annual Meeting of the National Council for Measurement in Education (Washington, D.C., March, 1975)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS *Criterion Referenced Tests; *Evaluation Methods; Higher Education; Reading Achievement; *Reading Development; Reading Improvement; *Reading Research; Reading Skills; *Test Construction

ABSTRACT

This study was designed to investigate the construction of a new measure for assessing reading performance which would prove to be a more sensitive and useful instrument for measuring the achievement growth induced by reading programs than the usual standardized tests. This report outlines the procedures followed in generating the model and constructing the tests, and presents some preliminary data collected in the course of field testing the new instrument. The hypothesis that the criterion-referenced tests would prove to be more sensitive measures of reading improvement over a lengthy period of time when compared to norm-referenced tests was partially confirmed. The results of the study are presented in narrative and table form. (RB)

A Criterion-Referenced Testing Model for Assessing Growth in Reading

Irene Athey
University of Rochester
Rochester, New York

Robert P. O'Reilly
State Education Department
Albany, New York

The current pressures for increased productivity and efficiency in education has led, in recent years, to the search for new evaluation procedures which are capable of measuring the outcomes of school instruction with more precision than traditional standardized tests. Specifically, the most urgent need is for measures which are relevant to a school system's program objectives, and which can provide information useful in making decisions about the merits of its programs.

This study was designed to investigate the possibility of constructing a new measure of reading performance which, when administered in a longitudinal matrix-sampling design, might prove to be a more sensitive and useful instrument for measuring the achievement growth induced by reading programs than the usual standardized tests. This report will outline the procedures followed in generating the model and constructing the tests, and will present some preliminary analyses of the data collected in the course of field-testing the new instrument.

The initial source of input to the test development process was provided by the Bank of Reading Objectives (BRO) compiled and organized over the period of a year by a team of reading researchers and curriculum experts under the auspices of the Bureau of School and Cultural Research of the New York State Education Department. During the course of this enterprise, many reading

---

Paper presented at the annual conference of the National Council for Measurement in Education, Washington, D. C., March 31, 1975.

systems were examined, but the final product of some 2000 objectives borrowed heavily from the SOBLR system developed by Skager at UCLA and Cohen's reading system. After considerable rewriting and modification, the Bank was organized into six areas: Multisensory Readiness, Decoding, Vocabulary, Comprehension, Location and Study Skills, and Reading in the Content Areas. In the initial phase of the study a decision was made to confine the construction of test items to two areas, Vocabulary and Comprehension. Committees of teachers and reading specialists in each of the nine participating school districts selected from the total bank those objectives in the areas of vocabulary and comprehension which seemed most pertinent to their own reading programs. In addition, they indicated the relative emphasis which the district placed on the skills represented by those objectives, assigning to each objective the number of test items they wished to have constructed. At the outset it was agreed that the total test should not exceed 30 minutes, so the committees were under some constraints to ensure that all important objectives were indeed adequately represented. Lists of basic sight words and of graded reading materials supplied by the school districts were used by the test construction team as a guide in the selection of passages and vocabulary items.

For the pilot phase, grades 4 through 6 were selected for intensive study. However, criterion-referenced tests must be geared to the level at which students are actually achieving, irrespective of their current grade placement. Since the range of achievement was approximately seven years for these three grades, seven levels of the tests, corresponding roughly to grades 1 through 7, were constructed.

The longitudinal design of the study called for repeated measurement on the same subjects at regular intervals determined by the school district. For example, a district might elect to administer the tests five times a year at

intervals of two months. If the district or the teachers were reluctant to retest so often, or at such short intervals, they had the option of confining the testing to three administrations at the beginning, middle, and end of the school year. However, in the pilot phase, all school districts administered the tests five times at two-weekly intervals between March and June, 1974. To enable them to do so, five equivalent forms of the tests were constructed at each level, and students were assigned in random order to the various forms, in such a way that ultimately every student took all five forms of the test at the level to which he had been assigned by his teacher. Approximately 4000 students constituted the sample, and they came from school districts with a wide range of demographic characteristics, including urban centers in the largest cities of New York State, and some suburban and rural areas.

Concurrent with the testing program, a technological support system was developed at a central location and implemented at various locations adjacent to the participating school districts. The tests were printed and assembled at the central location and shipped directly to the schools. Following each test administration, they were scored at the nearest computer facility, and copies of the data analyses returned to the central location and to the school district. Feedback to the schools consisted of individual and group scores on every item, every objective, and the total test. Using this information, the teacher could ascertain the effectiveness of instruction on any specific skill from one test administration to the next, or even the extent of students' mastery of a skill prior to instruction, for example, at the beginning of the school year. The data could thus serve not only to diagnose the strengths and weaknesses of particular students, but to suggest more effective uses of the available instructional time both for individual and group purposes. Similar feedback was also provided to each student in the

form of a coupon showing his progress on each item and skill and on the total test for each administration. Teachers have reported informally that provision of this type of feedback in which students monitor their own progress can be highly motivating, and this assertion may be tested empirically in a later phase of the study.

An important objective of the pilot study was to obtain feedback from the schools on both the test instrument and the computer support system. A questionnaire was distributed to all cooperating teachers, in which they were encouraged to critique the format and composition of the tests, and to comment on the usefulness of the feedback in the form in which it was presented. Several major modifications and refinements were introduced as a result of this input from the schools. First, it became apparent that comprehension was the overriding concept, and that most of the skills which had been designated as vocabulary objectives could readily be subsumed under comprehension. At this time an attempt was also made to reconceptualize the notion of reading comprehension to determine whether all the major parameters had been included in the operational definition constituted by the objectives which had been consensually validated by the reading teachers. Every school district, for example, recognized the importance of certain cognitive skills such as the ability to classify or to designate cause and effect. On the other hand, few included the ability to process complex syntactic structures, although such an ability must clearly be related to the comprehension of printed text.* Second, it was found that even though the tests had been based on graded instructional materials and trade books, some of the passages and items were too difficult for the levels at which they had been written. All passages used in the second round of the study were therefore submitted to a Dale-Chall readability formula, with additional checks

*For further discussion of this topic see Athey (1975).

for vocabulary level in the Harris-Jacobson Basic Word Lists. In addition, the questions were carefully screened to ensure that the passages must be read in order to obtain the necessary information for answering the questions, and to eliminate "keying" of one item by another. An attempt was also made to sample the universe of "real life" reading materials in a way that would adequately represent the various domains of interest for students of a given age range. Third, the original seven levels were expanded to 20 levels, two per grade for grades 1 through 10. This extension permits a school district either to test at more frequent intervals or to enjoy greater leeway in the selection of items in the compilation of their own tests. To facilitate this selection, the test items are currently presented in the form of a Test Development Notebook consisting of 800 pages containing some 4000 items. The items are grouped in about 20 content categories representing skills of literal and interpretive comprehension. Thus, a typical page of the notebook might contain a passage of the length prescribed for a particular level and a maximum of five items representing two or three objectives. These pages can be removed from the notebook to be reproduced and assembled by the school. This arrangement gives the school district considerable flexibility in assembling its own test packages without affecting the overall research design of the study.

The major hypothesis of the longitudinal study was that the criterion-referenced tests would prove to be more sensitive measures of growth in reading achievement over a long-term period than norm-referenced standardized tests. A context for examining the issue of test sensitivity was created in another phase of the study by gathering data on teacher and student characteristics and on the quantity and quality of the instructional process. Student characteristics data included individual and group socioeconomic status and

achievement and/or ability scores available from previous years of schooling, Instructional process data were gathered through individual interviews with teachers, specialists, and aides involved in the classroom and in special reading programs. The quantity and quality of instruction available to each student in each program condition were quantified. The data set included estimates of the number of instructional minutes available to individual students under different resource conditions, personnel configurations, materials, equipment, facilities, teacher experience and training, and classroom organization (individual, small group, or large group). The notion guiding the collection of student and teacher characteristics data was that factors under the control of the school would provide appropriate criteria for differentiating between achievement tests on the issue of sensitivity. Specifically, it was hypothesized that school factors should contribute relatively more strongly to changes in scores on the criterion-referenced tests.

In examining the first results of the analyses designed to test this hypothesis, it should be borne in mind that the data are those gathered in the pilot phase, before the criterion-referenced tests underwent considerable modification. The results should therefore be considered as highly tentative, and interpretations or implications should be drawn in the light of that fact.

Two analyses which were available at the time of writing, using the CRT and the California Achievement Test (Reading, Total Score) as criteria in the regression equations at grade levels 4 and 5 are summarized in Tables 1 and 2. At both levels, the CRT proved to have substantial predictive value, with the percentage of white students in the class, the Grade 3 PEP scores, the amount of individual help given by the teacher, and the number of pupils in the class being other significant predictors. The multiple correlations of the 22

Table 1

Comparison of Predictive Variables Using Criterion-Referenced Test and California Achievement Test as Criteria

Level 4 (N=607)

Independent Variables	r*	wt.
Criterion-Referenced Test		
CRT pretest	.57	.41
Percentage white students in class	.32	.29
Grade 3 PEP scores	.31	.22
$r_{c.1...22}$	=.68	
California Achievement Test		
CRT pretest	.49	.33
Percentage of white students in class	.47	.31
Grade 3 PEP scores	.37	.29
$r_{c.1...22}$	=.68	

*A correlation of .09 is significant at the .05 level for N=500.

Table 2

Comparison of Predictive Variables Using Criterion-Referenced Test and California Achievement Test as Criteria

Level 5 (N=497)

Independent Variables	r*	wt.
Criterion-Referenced Test		
CRT pretest	.69	.49
Minutes per year of individual help by teacher	.28	.13
$r_{c.1...22}$	=.76	
California Achievement Test		
CRT pretest	.37	.39
Number of pupils in class	.22	.26
Grade 3 PEP scores	.37	.23
$r_{c.1...22}$	=.63	

*A correlation of .09 is significant at the .05 level for N=500.

independent variables to the two criteria are also shown. At the fourth grade level, there was no difference in the sensitivity of the CRT and the CAT to the total process and characteristics variables, but at the fifth grade level, the correlation for the CRT criterion was substantially higher (r=.76) than for the CAT (r=.63). This suggests that the initial hypothesis has some plausibility, and we would expect, with further refinement of the tests, that the results would be even stronger in this direction. We would also hope that variables outside the instructional process (such as socioeconomic status) would become less evident, so that the criterion tests become purer measures of the outcome of the instructional process. The present study has pointed the direction and provided results which are sufficiently encouraging to suggest that our original hypotheses were well-founded.

References

Athey, I. Children's understanding of syntax in relation to reading comprehension. Paper presented at the annual conference of the International Reading Association, New York, May 1975.